Milestone One — Auto Insurance Fraud Detection

Student: Violette Similien Volodkevich  
Course: DX799 O1 Data Science Capstone (Fall 2025)  
Date: October 20, 2025

# 1. Problem Statement

Auto insurance fraud imposes substantial costs on policyholders and insurers. This project develops predictive models that flag likely fraudulent claims using structured claim, policy, and incident data from the carclaims 12.csv dataset. The target variable is FraudFound (1 = fraud, 0 = non-fraud). The goal is to create interpretable and scalable models to reduce fraudulent payouts and improve claim accuracy.

# 2. Data and Preparation

The dataset contains both numerical and categorical features such as claim details, vehicle characteristics, and policy information. Missing numerical values were imputed with the median, while missing categorical values were imputed with the mode. Categorical features were one-hot encoded and numerical features were standardized for models sensitive to scale. An 80/20 stratified train-test split was used to preserve class balance for FraudFound.

# 3. Modeling Summaries (Weeks 1–6)

## Week 1 — Polynomial & Interaction Terms

Polynomial (degree=2) and interaction features modeled non-linear relationships. A logistic regression model was trained on these engineered features. Multicollinearity checked via VIF.

[Insert ROC, PR, Confusion Matrix plots here]

## Week 2 — Regularization (Ridge, Lasso, Elastic Net)

L1, L2, and Elastic Net regularization compared to reduce overfitting. Hyperparameters tuned via 5-fold CV, best model selected by ROC-AUC.

[Insert plots here]

## Week 3 — Feature Selection & Dimensionality Reduction

Forward/backward selection identified key predictors. PCA retained ~90% variance, followed by logistic regression on components.

[Insert plots here]

## Week 4 — Logistic Regression & Feature Scaling

Baseline classification using scaled numerics and encoded categoricals. Penalty and regularization tuned for optimal bias-variance tradeoff.

[Insert plots here]

## Week 5 — Support Vector Machines (Kernels & Regularization)

Linear and RBF kernels compared. Regularization (C) tuned by grid search; RBF slightly outperformed linear.

[Insert plots here]

## Week 6 — Decision Trees & Random Forests

Decision Trees produced interpretable rules; Random Forests reduced overfitting through ensembling. RF achieved the best overall AUC.

[Insert ROC, PR, Confusion Matrix, Feature Importance plots here]

# 4. Deep Dive — Logistic Regression and Random Forest

The deep dive focused on Logistic Regression (Week 4) and Random Forest (Week 6). Logistic Regression offered transparency and interpretability, while Random Forest captured non-linear interactions. Both tuned using 5-fold CV. Random Forest achieved highest ROC-AUC; Logistic Regression remained an interpretable baseline.

# 5. Overfitting and Hyperparameter Tuning

Overfitting controlled via cross-validation, regularization, and tree pruning. Parameters (C, α, depth) tuned using grid search. Final models validated on hold-out test set for generalization.

# 6. Evaluation Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | F1 Score | ROC-AUC | PR-AUC |
| Week 1 | [ ] | [ ] | [ ] | [ ] |
| Week 2 | [ ] | [ ] | [ ] | [ ] |
| Week 3 | [ ] | [ ] | [ ] | [ ] |
| Week 4 | [ ] | [ ] | [ ] | [ ] |
| Week 5 | [ ] | [ ] | [ ] | [ ] |
| Week 6 | [ ] | [ ] | [ ] | [ ] |

# 7. Expected vs. Unexpected Findings & Role of EDA

EDA revealed strong correlations between incident severity, policy details, and fraud likelihood. Expected: higher fraud risk in single-vehicle accidents without police reports. Unexpected: some demographic groups showed elevated false positives.

# 8. Conclusion and Next Steps

Random Forest delivered the best discrimination power, while Logistic Regression offered interpretability. Next steps include addressing class imbalance, applying SHAP for explainability, and integrating vehicle image data for hybrid detection.

# References

* Aqqad, A. (2023). Insurance\_claims. Mendeley Data. https://data.mendeley.com/datasets/992mh7dk9y/2
* Kapoor, K. (2023). Vehicle Insurance Fraud Detection. Kaggle. https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection
* Humans In The Loop. (2023). Car Parts and Car Damages. Kaggle. https://www.kaggle.com/datasets/humansintheloop/car-parts-and-car-damages
* FBI. (2010). Insurance Fraud. FBI. https://www.fbi.gov/stats-services/publications/insurance-fraud
* Dell, E. (2024). How virtual inspections can benefit insurance customers. Digital Insurance. https://www.dig-in.com/news/how-virtual-inspections-can-benefit-insurance-customers

# Appendix — Plots and Visuals

Insert saved plots from each week's notebook (ROC, PR, Confusion Matrix, Feature Importance).